Tag-Analysen

Statistik und Hypothesen

Das Ziel ist eine allgemeine Untersuchungsmethodik zu entwickeln, die auf verschiedene Tagging-Systeme angewandt werden kann.

Dimensionen der Tag-Analyse

Tags können empirisch entlang der Dimensionen „Form“, „Funktion“ und „Herkunft“ analysiert werden.

Formanalyse

- **Wortanzahl**
  - einzeln (mehrheitlich)
  - mehrere, Leerzeichentrennung
    - Substantive (mehrheitlich)
    - Substantiv + andere
    - andere
  - Wortketten (mit Unterstrich)
    - Substantive
    - Substantiv + andere
    - andere

- **Wortart**
  - Substantive
  - Adjektive
  - (Ad-) Verben
  - Akronyme
  - Zahlen und andere Zeichen

- **Rechtschreibung**
  - korrekt
  - fehlerhaft
  - Varianten

- **Personalisierungen**
  - Neologismen
  - Eigenschöpfungen
  - Neukompositionen
  - Funktionswörter

- **Sprache**
  - Deutsch
  - Englisch
  - Andere

Funktionsanalyse

- **inhaltsorientiert**
    - formal
        - Autor
        - Publikationsform
        - Datum
        - Herkunft
        - Sprache
    - inhaltlich
        - direkte Beschreibung
        - Fach- /Themenbereich
        - Klassifikationsversuch
        - Kategorie
            - Bericht
            - Besprechung
            - Anleitung
            - ...
        - Methodik
            - empirisch
            - theoretisch
            - ...
- **nicht inhaltsorientiert** (durch klassische Sacherschließung nicht abgedeckt)
    - emotional
        - positiv
        - negativ
    - Zeit- und Arbeitsplanung
        - Aufgaben
        - Abläufe
        - Kontext
        - Markierungen
            - exklusiv
            - geteilt
            - "wichtig" usw.
    - Vermeidung (mehrheitlich)

Herkunftsanalyse

- **aus dem Volltext**
    - aus dem Titel (1/2)
    - aus dem Abstract
    - aus dem Volltext
    - aus den Schlagworten
- **Variation des Volltextes**
    - Rechtschreibfehler
    - flektierte Formen

        - Wortstämme
        - aus dem Titel: 63%

- **nicht im Volltext (1/3)**
    - Synonyme
    - Hyponyme (Unterbegriff)
    - Hyperonyme (Oberbegriff)

- **Tagger**
    - Nutzer
        - alle
            - weniger und einfachere Konzepte
            - Vermeidung sehr spezifischer Begriffe
            - allgemeinere Kategorien
            - kürzere Worte
        - eingeloggte
    - Experten
        - Fachreferenten
        - Autoren
            - mehr Worte pro Schlagwort
            - 2/3 der Schlagworte werden von Nutzern nicht verwendet

Tag-Analysen im Sinne des CT-Projekts

In der ersten Untersuchungsphase können wir auf zwei verschiedene Datensätze zurückgreifen. Wir trennen zwischen den Tags der Nutzer der UB Mannheim und jenen aus den übrigen Quellen. Insofern gilt für die folgenden Analysen im Rahmen der Zielsetzung des CT-Projekts zusätzlich immer eine direkte Vergleichsmöglichkeit zwischen den Datensätzen.

„Allgemeine“ Statistik

1. Absolute Häufigkeiten: Tags pro Titel, User und im Gesamtdatensatz

   Erkenntnisziele: Erschließungsgrad der Titel, Aktivitätsprofil der User, inhaltliche Ausrichtung von Nutzern[1], Nutzergruppen bzw. aller Nutzer

2. Dynamische Häufigkeiten: Tags im Zeitverlauf (einzeln, pro Titel, pro Nutzer)

   Erkenntnisziele: „Suggestive Tagging“ (Kopieren, Imitieren), „natürliche“ Tendenz zur Konvergenz, Nutzerverhalten über die Zeit bei Intensivnutzern

3. POS-Häufigkeiten: Wortarten, Akronyme und andere Zeichen

   Erkenntnisziele: Qualitätsanalyse, Anteil „wohl geformter“ Tags

4. Wortlänge

   Erkenntnisziele: Komplexität und Spezialisierung der Erschließung

1 Nur sinnvoll bei entsprechender Fallzahl.

(Semi-) Qualitative Verfahren

1. Clustering: Nutzergruppen, Quellengruppen, Tag-Gruppen

Erkenntnisziele: Community Building, Quellengruppen aus „Laiensicht“, Ähnlichkeiten von Tags zur Synonymvermutung

2. Klassifikationsmapping: Zuordnung frei vergebener Tags zu bestehenden Klassifikationen (über MABXML?!)

Erkenntnisziele: Nachnutzbarkeit von CT im Rahmen professioneller Sacherschließung, Möglichkeiten automatisierter Hierarchisierung von Annotationen

3. exemplarischer Direktvergleich zwischen CT und SWD

Erkenntnisziele: Innovationscharakter des CT, Überprüfung gängiger Annahmen bezüglich der Mängel sowie der Vorteile von CT-Systemen

Hypothesen

Aus den geschilderten Analysemöglichkeiten sowie der hierzu bereits erschienenen Literatur lassen sich z.B. die folgenden Hypothesen ableiten und überprüfen:

Quantitäten:

1. Es gibt wenige Nutzer, die viele Tags für viele Publikationen vergeben sowie viele Nutzer, die wenig aktiv sind.
2. Die meisten Titel werden mit wenigen oder gar keinen Tags („imported“ etc.) erschlossen.
3. Wenige Tags werden häufig, viele Tags selten benutzt („Power-Law“)
4. Der Erschließungsgrad von Quellen, die informationstechnologischen Fächern zugeordnet werden können, ist wesentlich höher, als jener anderer Quellen.
5. Nutzergruppen bleiben ihren Themen treu.
6. Der „Priming Effekt“ durch vorhandene Tags für eine Quelle ist empirisch nachweisbar („Wer hat, dem wird gegeben“).
7. Tags haben überwiegend die folgenden Eigenschaften: sie stammen direkt oder als leichte Variation aus dem Titel, sie sind Substantive und sie stehen meistens als einzelne Wörter.

Qualität:

1. Der Anteil falsch geschriebener Begriffe ist gering.
2. Die Tags sind zumeist unspezifisch, kurz und eher banal.
3. Mit zunehmender Anzahl verschiedener Tags für eine Quelle sinkt dessen inhaltliche Diskriminierung (Präzision der Erschließung).
4. Tags eignen sich zur Exploration und zur Assoziation, jedoch nicht zur Optimierung einer Treffermenge nach konkreten Suchzielen. Deshalb können sie insbesondere zur Pflege bestehender Klassifikationen beitragen (Aktualität, Umfang).
5. Eine erhebliche Anzahl von Tags wird soweit vorhanden aus bereits vorhandenen Schlagwörtern übernommen.
6. Je häufiger ein Tag im Datensatz vorkommt, desto allgemeiner ist sein inhaltlicher Gehalt.